CELEBAL TECHNOLOGIES PVT. LTD.

PROJECT

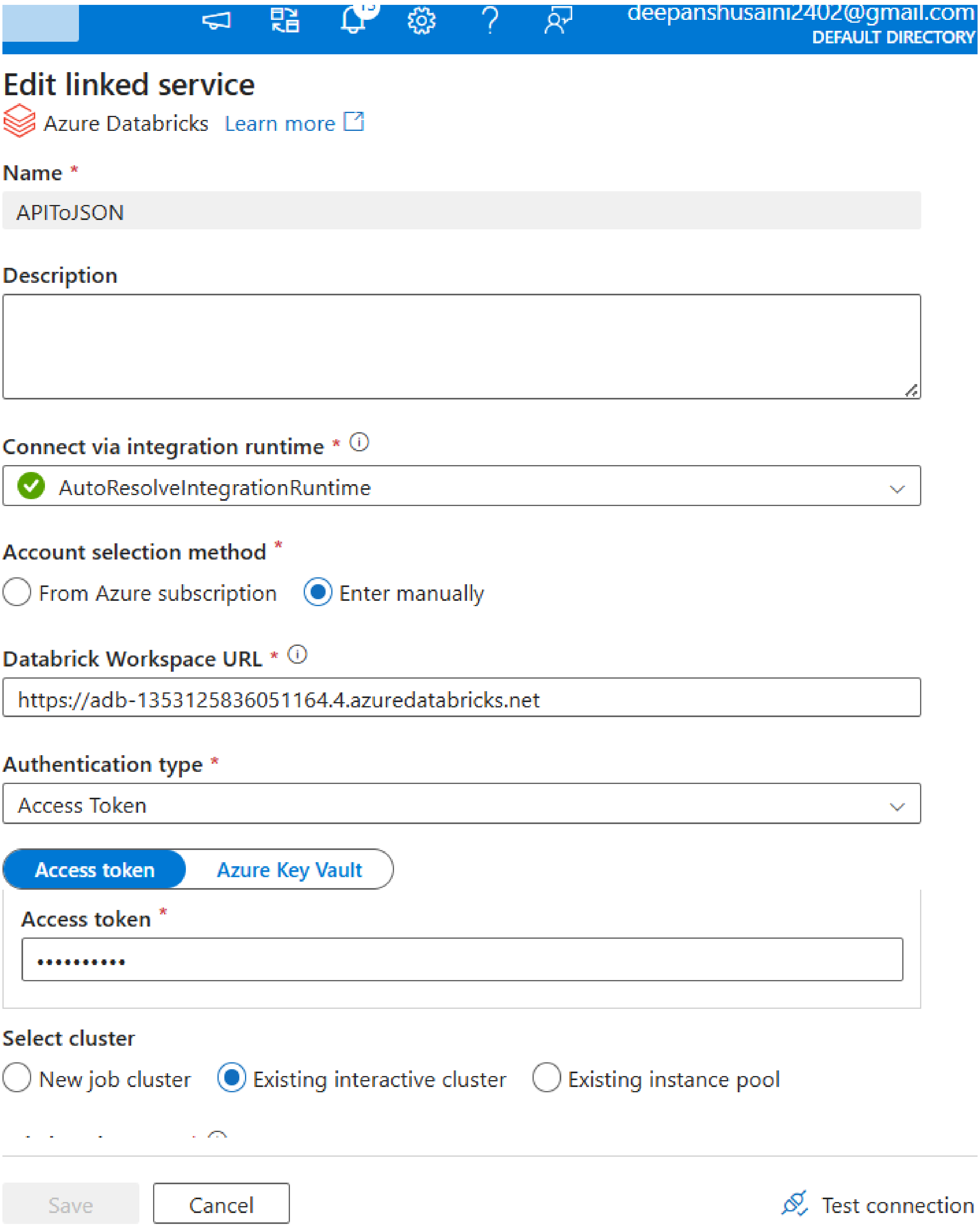
PROBLEM STATEMENT: Generic Continuous data Ingestion from multiple streaming sources into databricks.

1. Create a pipeline to fetch the 5 countries (india,us,uk,china,russia) data from Rest API (https://restcountries.com/v3.1/name/{name} here replace the {name} with Country name like https://restcountries.com/v3.1/name/us) and save it in separate file as JSON with File name equal to Country name.

STEP 1: Login to Azure Portal and create an Azure Databricks service.

STEP 2: In Azure Databricks, create notebook apiToJson.

STEP 3: Launch Azure Data Factory and create a linked service (APIToJSON) linking databricks with data factory.



Edit linked service

Azure Databricks Learn more

Name *

APIToJSON

Description

Connect via integration runtime *

✓ AutoResolveIntegrationRuntime

Account selection method *

○ From Azure subscription ● Enter manually

Databrick Workspace URL *

https://adb-1353125836051164.4.azuredatabricks.net

Authentication type *

Access Token

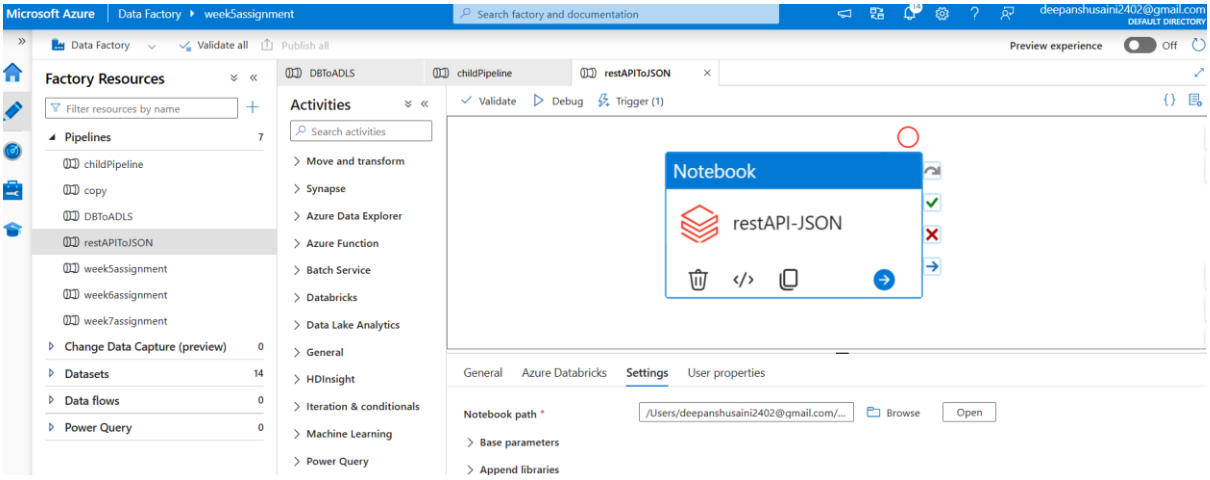
Access token *

••••••••••

Select cluster

○ New job cluster ● Existing interactive cluster ○ Existing instance pool

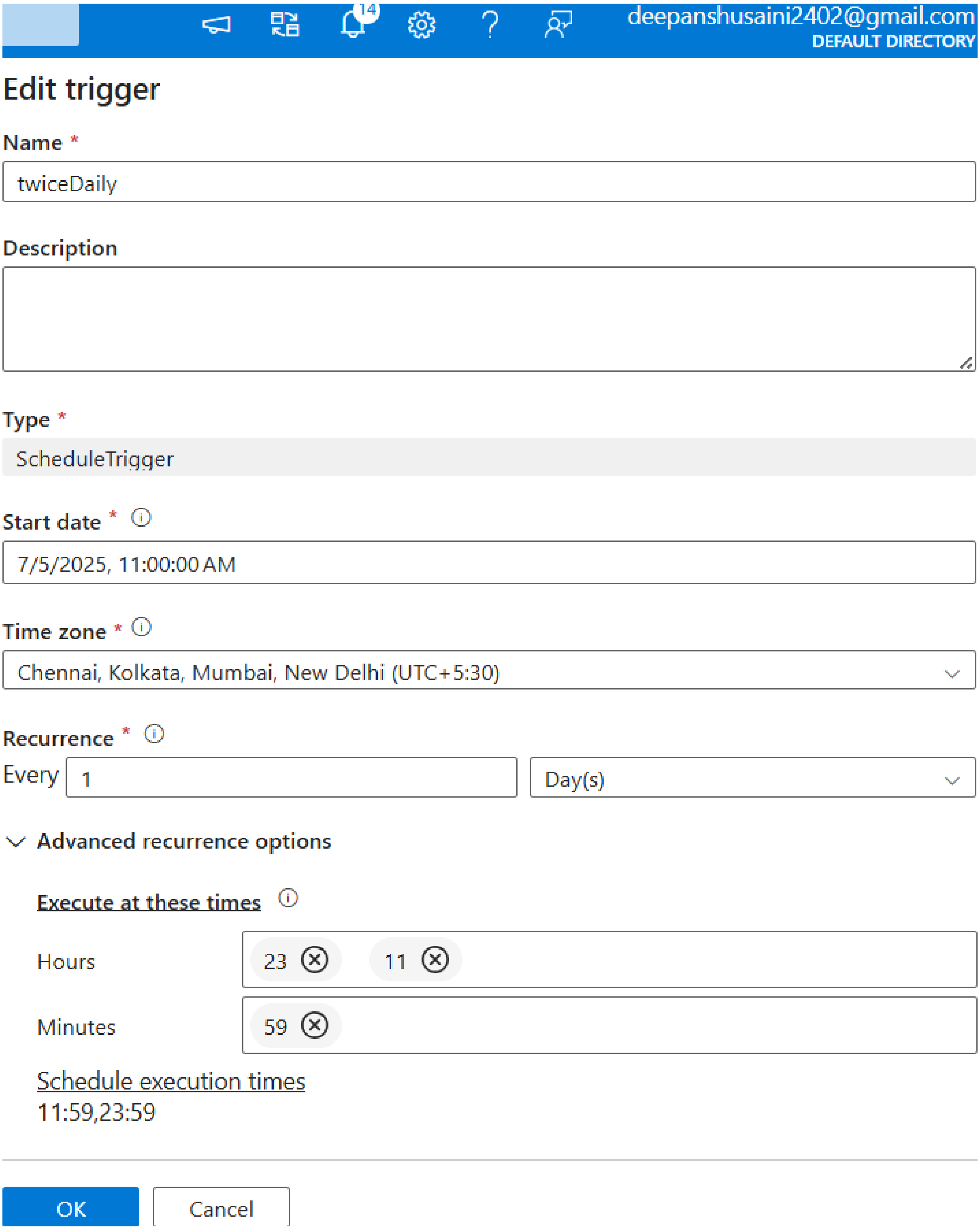
STEP 4: Create pipeline (restAPIToJSON) with a databricks Notebook Activity (restAPI-JSON) and specify path for apiToJson notebook in databricks.



Notebook path * /Users/deepanshusaini2402@gmail.com/...

2. Add the trigger to above pipeline in such a way that it will automatically run two times in a day (12:00 AM and 12:00 PM IST).

STEP 1: Add a Scheduled Trigger (twiceDaily).



deepanshusaini2402@gmail.com
DEFAULT DIRECTORY

Edit trigger

Name *

twiceDaily

Description

Type *

ScheduleTrigger

Start date *

7/5/2025, 11:00:00 AM

Time zone *

Chennai, Kolkata, Mumbai, New Delhi (UTC+5:30)

Recurrence *

Every 1 Day(s)

Advanced recurrence options

Execute at these times

Hours 23 11

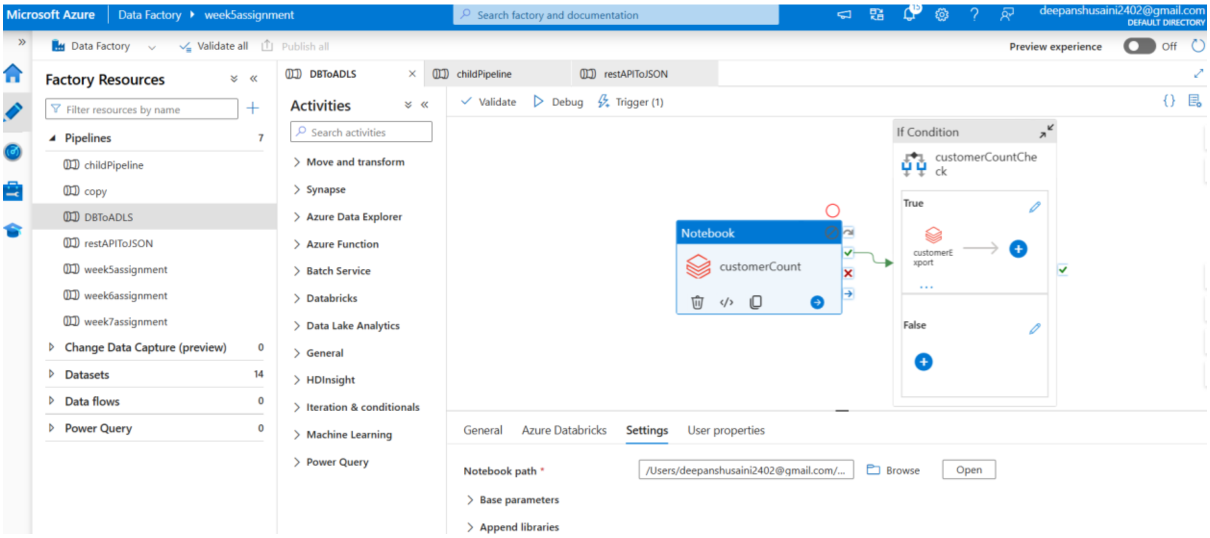
Minutes 59

Schedule execution times
11:59,23:59

3. Create a pipeline to copy customer data from db to adls only if record count is more than 500. Once data gets copied it should call a child pipeline (which will copy the product data from table if customer record count is > 600).

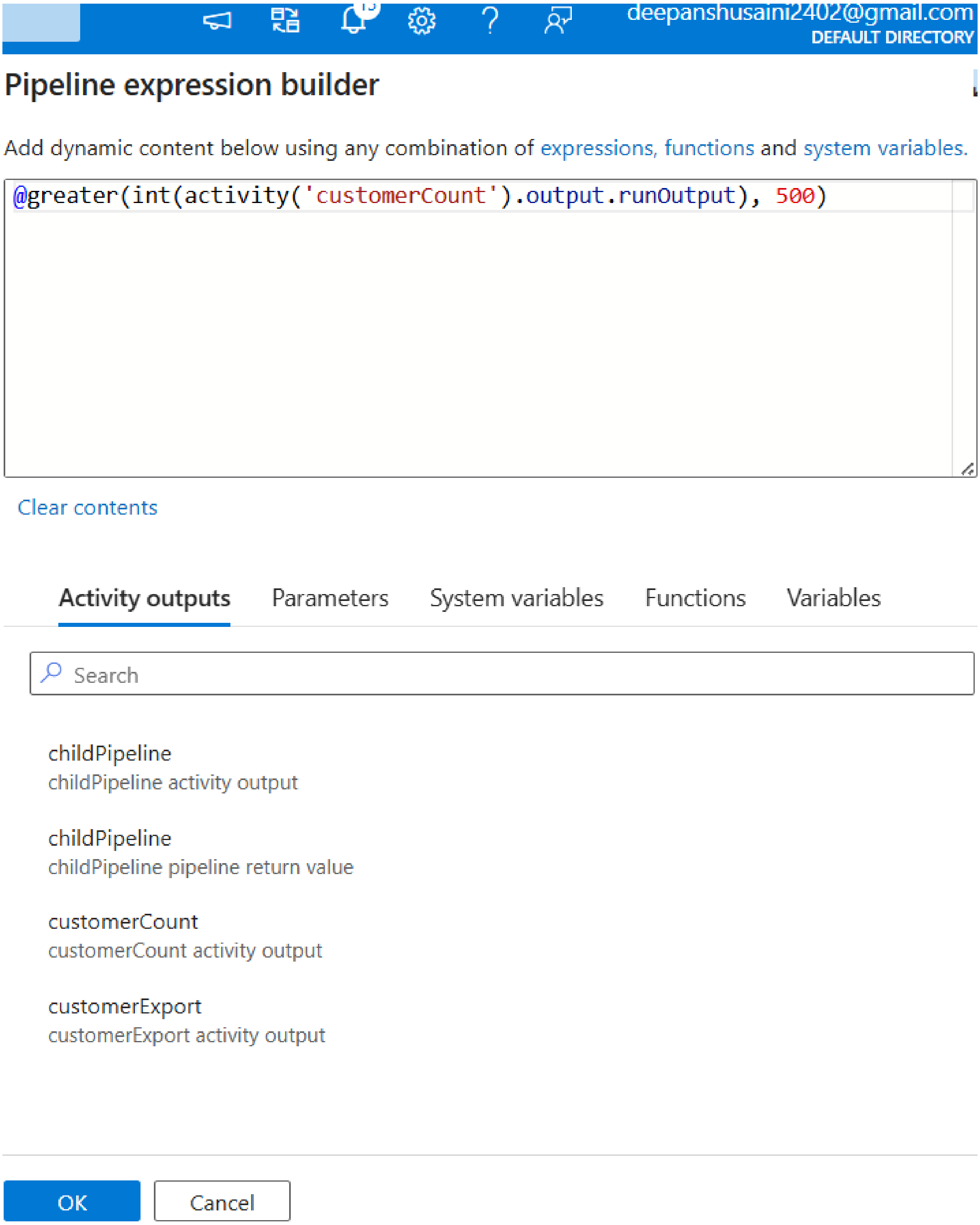
STEP 1: In databricks create a notebook (dbToADLS) to create databases in databricks named CustomerData and ProductData.

STEP 2: Create pipeline (DBToADLS) and add databricks Notebook Activity (customerCount) by specifying path for customerCount notebook in databricks to get the customer count from Customer Data table.



Notebook path * /Users/deepanshusaini2402@gmail.com/...

STEP 3: Add If Condition Activity (customerCountCheck) with below given expression.



Pipeline expression builder

Add dynamic content below using any combination of expressions, functions and system variables.

```
@greater(int(activity('customerCount').output.runOutput), 500)
```

childPipeline
childPipeline activity output

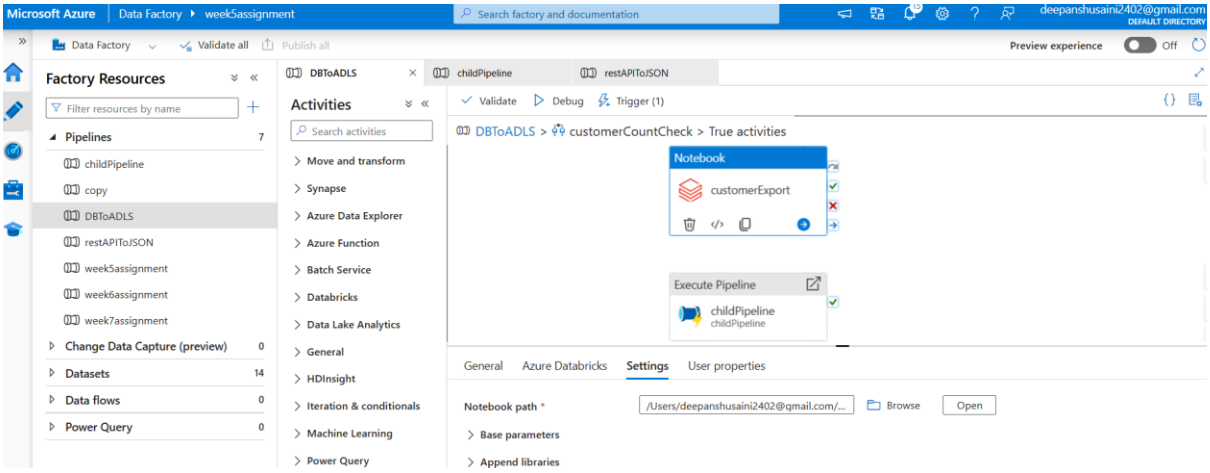
childPipeline
childPipeline pipeline return value

customerCount
customerCount activity output

customerExport
customerExport activity output

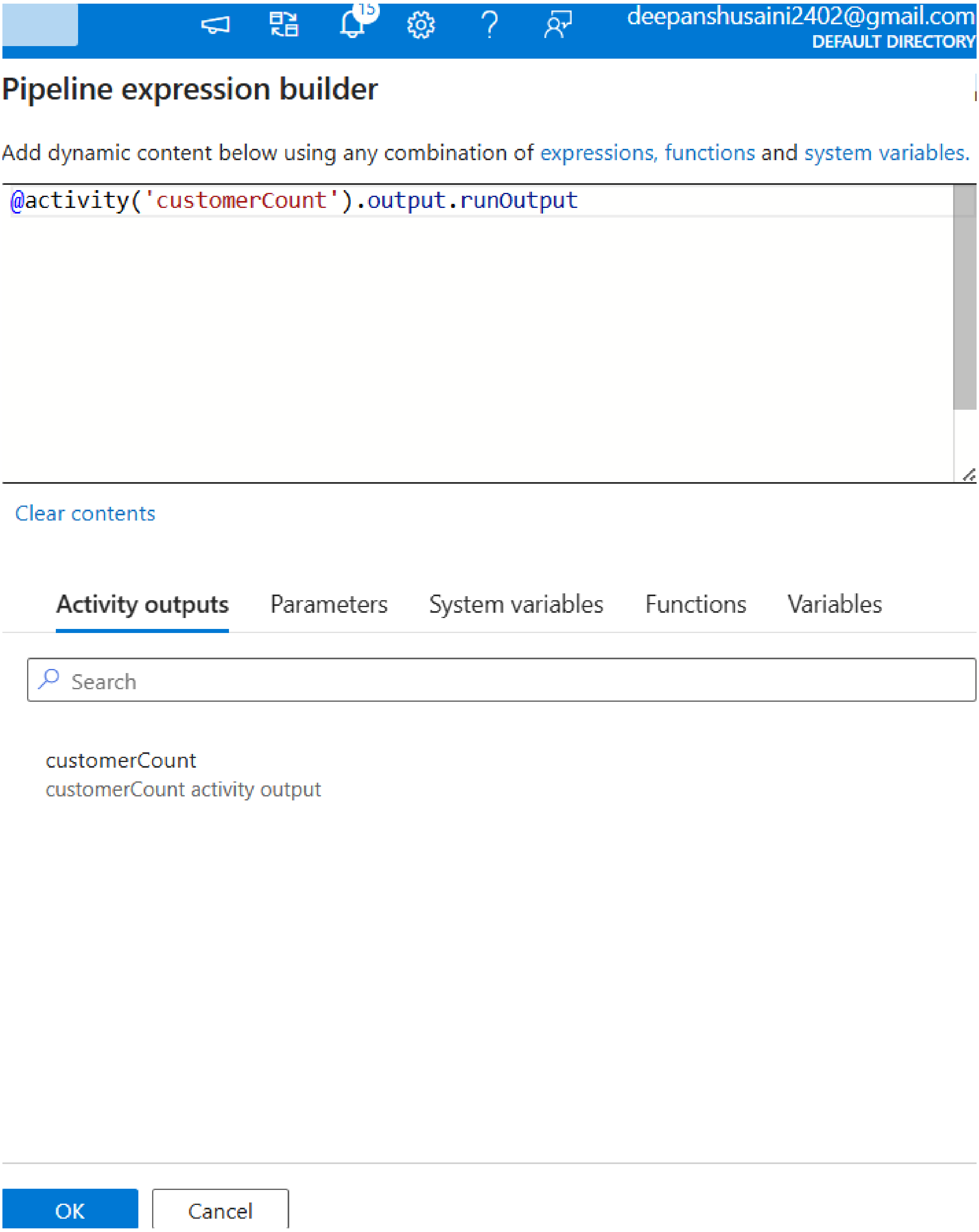
STEP 4: Inside True Case,

- Add databricks Notebook Activity (customerExport) linking it with customerExport databricks notebook to copy CuctomerData if count>500



Notebook path * /Users/deepanshusaini2402@gmail.com/...

- Add Execute Pipeline Activity (childPipeline) with a string type parameter named ‘customerCount’ with the below given value.



Pipeline expression builder

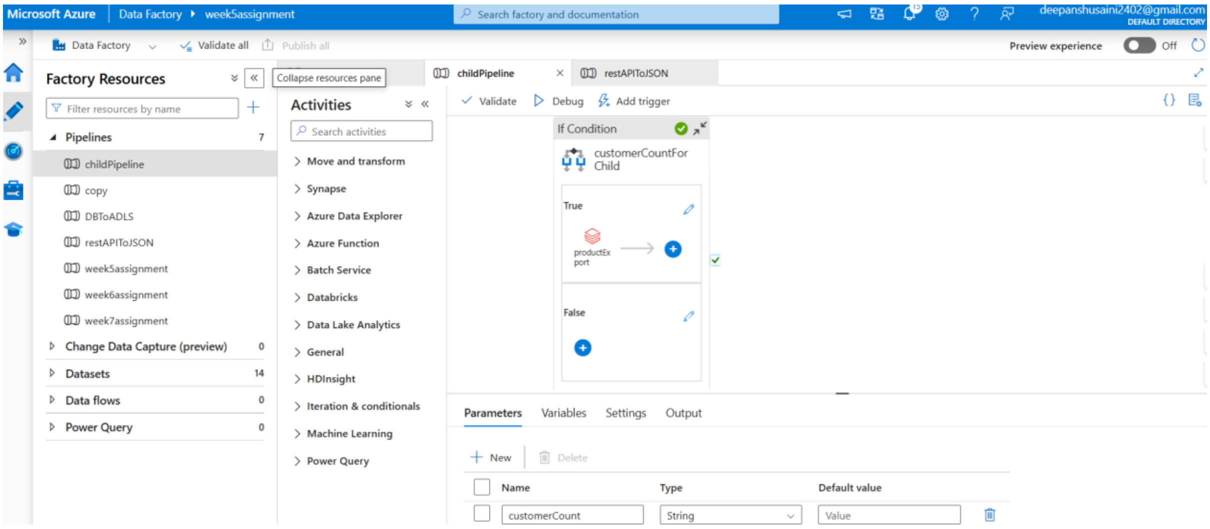
Add dynamic content below using any combination of expressions, functions and system variables.

```
@activity('customerCount').output.runOutput
```

customerCount
customerCount activity output

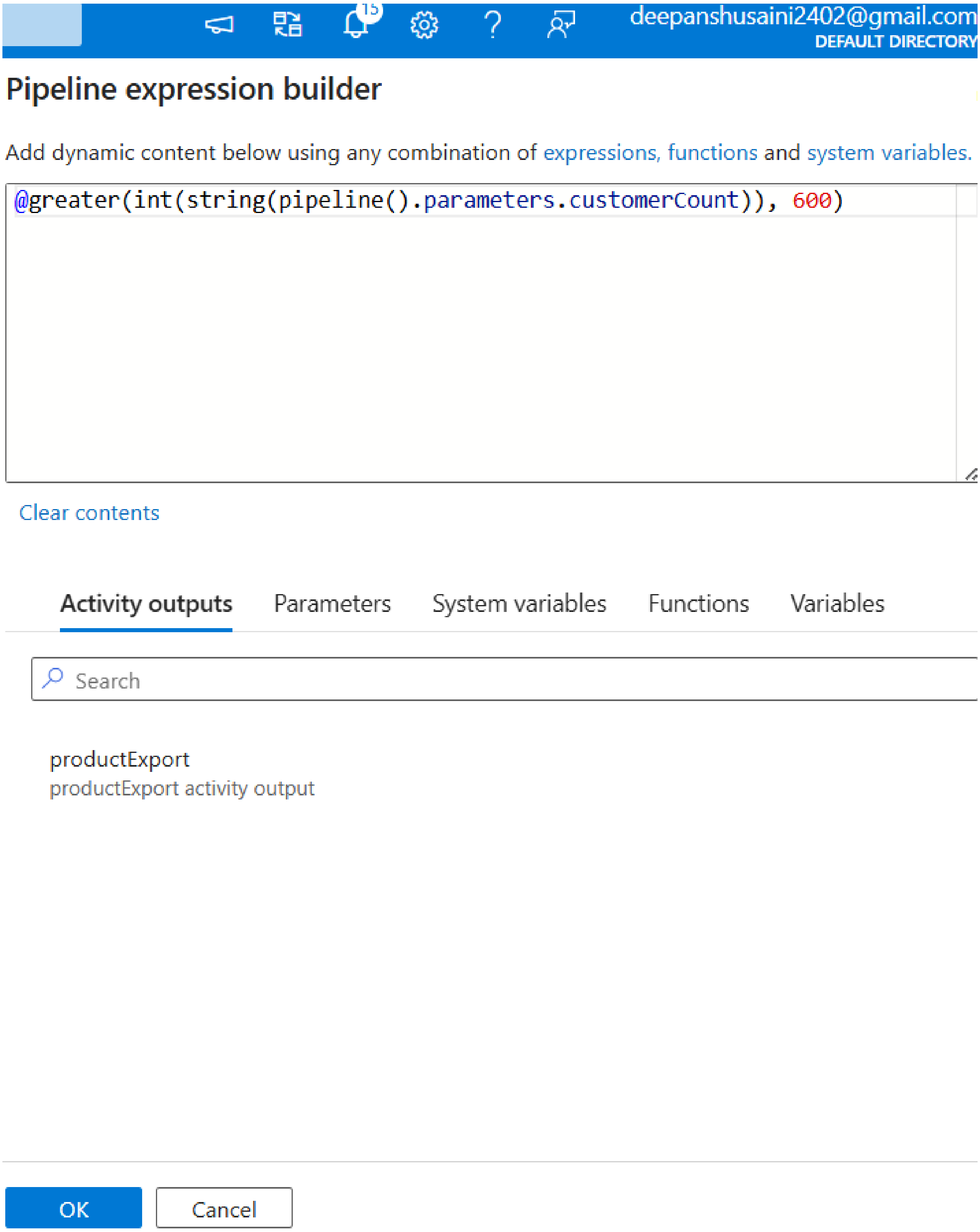
4. Design the pipeline in such a manner that the Customer pipeline will pass the customer count to the child product pipeline via Pipeline parameter.

STEP 1: Now create another pipeline (childPipeline) with string type parameter customerCount.



☐	Name	Type	Default value
☐	customerCount	String	Value

STEP 2: Add If Condition Activity (customerCountForChild) in it with following expression.

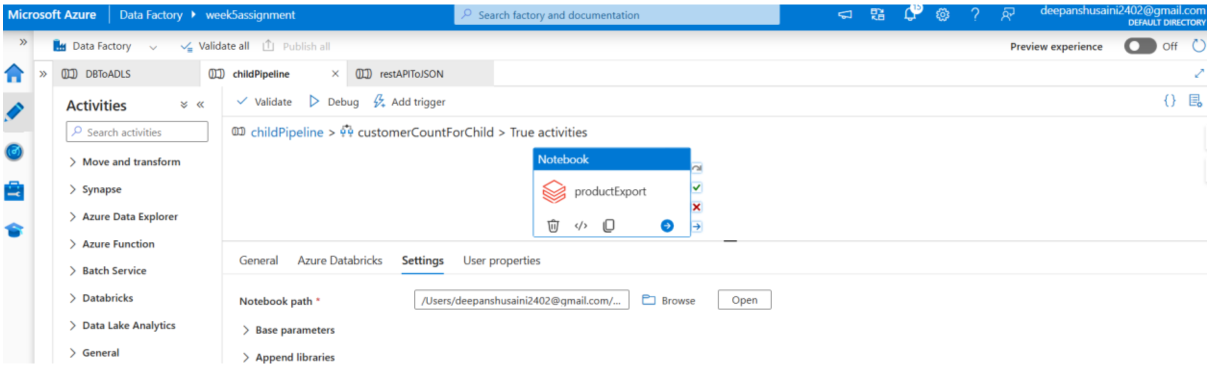


Pipeline expression builder

Add dynamic content below using any combination of expressions, functions and system variables.

```
@greater(int(string(pipeline().parameters.customerCount)), 600)
```

STEP 3: Inside True Case, add databricks Notebook Activity (productExport) linking it with productExport databricks notebook to copy productData to ADLS if customerCount>600.



childPipeline > customerCountForChild > True activities

Notebook: productExport

Notebook path * /Users/deepanshusaini2402@gmail.com/...